JMLR: Workshop and Conference Proceedings vol 35:1–18, 2014

Efficiency of conformalized ridge regression

Evgeny Burnaev EVGENY.BURNAEV@DATADVANCE.NET
Datadvance LLC, Institute for Information Transmission Problems, and Moscow Institute of Physics and Technology, Russia

Vladimir Vovk V.VOVK@RHUL.AC.UK
Computer Learning Research Centre, Department of Computer Science, Royal Holloway, University of London, United Kingdom

Abstract

Conformal prediction is a method of producing prediction sets that can be applied on top of a wide range of prediction algorithms. The method has a guaranteed coverage probability under the standard IID assumption regardless of whether the assumptions (often considerably more restrictive) of the underlying algorithm are satisfied. However, for the method to be really useful it is desirable that in the case where the assumptions of the underlying algorithm are satisfied, the conformal predictor loses little in efficiency as compared with the underlying algorithm (whereas being a conformal predictor, it has the stronger guarantee of validity). In this paper we explore the degree to which this additional requirement of efficiency is satisfied in the case of Bayesian ridge regression; we find that asymptotically conformal prediction sets differ little from ridge regression prediction intervals when the standard Bayesian assumptions are satisfied.

Keywords: Bayesian learning, conformal prediction, asymptotic analysis, Bahadur representation

1. Introduction

This paper discusses theoretical properties of the procedure described in the abstract as applied to Bayesian ridge regression in the primal form. The procedure itself has been discussed earlier in the Bayesian context under the names of frequentizing (Wasserman 2011, Section 3) and de-Bayesing (Vovk et al. 2005, p. 101); in this paper, however, we prefer the name "conformalizing". The procedure has also been studied empirically (see, e.g., Vovk et al. 2005, Figures 10.1–10.5, and Wasserman 2011, Figure 1, corrected in Vovk 2013b, Figure 11.1). To our knowledge, this paper is the first to explore the procedure theoretically.

The purpose of conformalizing is to make prediction algorithms, first of all Bayesian algorithms, valid under the assumption that the observations are generated independently from the same probability measure; we will refer to this assumption as the IID assumption. This is obviously a desirable step provided that we do not lose much if the assumptions of the original algorithm happen to be satisfied (metaphorically speaking, our method should do well even on its competitors' home turf). The situation here resembles that in nonparametric hypothesis testing (see, e.g., Randles et al. 2004), where nonparametric analogues of some classical parametric tests relying on Gaussian assumptions turned out to be surprisingly efficient even when the Gaussian assumptions are satisfied.

We start the main part of the paper from Section 2, in which we define the ridge regression procedure and the corresponding prediction intervals in a Bayesian setting involving strong Gaussian assumptions. It contains standard material and so no proofs. The following section, Section 3,

applies the conformalizing procedure to ridge regression in a way that facilitates theoretical analysis in the following sections; the resulting "conformalized ridge regression" is similar to but somewhat different from the algorithm called "ridge regression confidence machine" in Vovk et al. (2005).

Section 4 contains our main result. It shows that asymptotically we lose little when we conformalize ridge regression and the Gaussian assumptions are satisfied; namely, conformalizing changes the prediction interval by $O(n^{-1/2})$ with high probability, where n is the number of observations. Our main result gives precise asymptotic distributions for the differences between the left and right end-points of the prediction intervals output by the Bayesian and conformal predictors. These are theoretical counterparts of the preliminary empirical results obtained in Vovk et al. (2005) (Figures 10.1–10.5 and Section 8.5, pp. 205–207) and Vovk et al. (2009). We then discuss and interpret our main result using the notions of efficiency and conditional validity (introduced in the previous two sections). Section 5 gives a more explicit description of conformalized ridge regression, and in Section 6 we prove the main result.

Other recent theoretical work about efficiency and conditional validity of conformal predictors includes Lei and Wasserman's (2014). Whereas our predictor is obtained by conformalizing ridge regression, Lei and Wasserman's conformal predictor is specially crafted to achieve asymptotic efficiency and conditional validity. It is intuitively clear that whereas our algorithm is likely to produce reasonable results in practice (in situations where ridge regression produces reasonable results), Lei and Wasserman's algorithm is primarily of theoretical interest. A significant advantage of their algorithm, however, is that it is guaranteed to be asymptotically efficient and conditionally valid under their regularity assumptions, whereas our algorithm is guaranteed to be asymptotically efficient and conditionally valid only under the Gaussian assumptions.

2. Bayesian ridge regression

Much of the notation introduced in this section will be used throughout the paper. We are given a training sequence $(x_1, y_1), \dots, (x_{n-1}, y_{n-1})$ and a test object x_n, and our goal is to predict its label y_n. Each *observation* (x_i, y_i), $i = 1, \dots, n$ consists of an *object* $x_i \in \mathbb{R}^p$ and a *label* $y_i \in \mathbb{R}$. We are interested in the case where the number $n - 1$ of training observations is large, whereas the number p of attributes is fixed. Our setting is probabilistic; in particular, the observations are generated by a probability measure.

In this section we do not assume anything about the distribution of the objects $x_1, \dots, x_n$, but given the objects, the labels $y_1, \dots, y_n$ are generated by the rule

$$y_i = w \cdot x_i + \xi_i, \tag{1}$$

where w is a random vector distributed as $N(0, (\sigma^2/a)I)$ (the Gaussian distribution being parameterized by its mean and covariance matrix, and $I := I_p$ being the unit $p \times p$ matrix), each ξ_i is distributed as $N(0, \sigma^2)$, the random elements $w, \xi_1, \dots, \xi_n$ are independent (given the objects), and σ and a are given positive numbers.

The conditional distribution for the label y_n of the test object x_n given the training sequence and x_n is

$$N\left(\hat{y}_n, (1 + g_n)\sigma^2\right),$$

where

$$\hat{y}_n := x_n'(X'X + aI)^{-1}X'Y, \tag{2}$$

$$g_n := x_n'(X'X + aI)^{-1}x_n, \tag{3}$$

$X = X_{n-1}$ is the design matrix for the training sequence (the $(n-1) \times p$ matrix whose ith row is x_i', $i = 1, \ldots, n-1$), and $Y = Y_{n-1}$ is the vector $(y_1, \ldots, y_{n-1})'$ of the training labels; see, e.g., Vovk et al. (2005), (10.24). Therefore, the Bayesian prediction interval is

$$(B_*, B^*) := \left(\hat{y}_n - \sqrt{1+g_n}\sigma z_{\epsilon/2}, \hat{y}_n + \sqrt{1+g_n}\sigma z_{\epsilon/2}\right), \tag{4}$$

where ϵ is the significance level (the permitted probability of error, so that $1-\epsilon$ is the required coverage probability) and $z_{\epsilon/2}$ is the $(1-\epsilon/2)$-quantile of the standard normal distribution $N(0,1)$.

The prediction interval (4) enjoys several desiderata: it is unconditionally valid, in the sense that its error probability is equal to the given significance level ϵ; it is also valid conditionally on the training sequence and the test object x_n; finally, this prediction interval is the shortest possible conditionally valid interval. We will refer to the class of algorithms producing prediction intervals (4) (and depending on the parameters σ and a) as *Bayesian ridge regression* (BRR).

3. Conformalized ridge regression

Conformalized ridge regression (CRR) is a special case of conformal predictors; the latter are defined in, e.g., Vovk et al. (2005), Chapter 2, but we will reproduce the definition in our current context. First we define the *CRR conformity measure* A as the function that maps any finite sequence $(x_1, y_1), \ldots, (x_n, y_n)$ of observations of any length n to the sequence $(\alpha_1, \ldots, \alpha_n)$ of the following *conformity scores* α_i: for each $i = 1, \ldots, n$,

$$\alpha_i := \left|\{j = 1, \ldots, n \mid r_j \geq r_i\}\right| \wedge \left|\{j = 1, \ldots, n \mid r_j \leq r_i\}\right|,$$

where $(r_1, \ldots, r_n)'$ is the vector of ridge regression residuals $r_i := y_i - \hat{y}_i$,

$$\hat{y}_i := x_i'(X_n'X_n + aI)^{-1}X_n'Y_n$$

(cf. (2)), X_n is the overall design matrix (the $n \times p$ matrix whose ith row is x_i', $i = 1, \ldots, n$), and Y_n is the overall vector of labels (the vector in $\mathbb{R}^n$ whose ith element is y_i, $i = 1, \ldots, n$).

Remark 1 We interpret α_i as the degree to which the element (x_i, y_i) conforms to the full sequence $(x_1, y_1), \ldots, (x_n, y_n)$. Intuitively, (x_i, y_i) conforms to the sequence if its ridge regression residual is neither among the largest nor among the smallest. Instead of the simple residuals r_i we could have used deleted or studentized residuals (see, e.g., Vovk et al. 2005, pp. 34–35), but we choose the simplest definition, which makes calculations feasible. Another possibility is to use $-|r_i|$ as conformity scores; this choice leads to what was called "ridge regression confidence machines" in Vovk et al. (2005), Chapter 2, but its analysis is less feasible.

Given a significance level $\epsilon \in (0,1)$, a training sequence $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})$, and a test object x_n, *conformalized ridge regression* outputs the prediction set

$$\Gamma := \{y \mid p^y > \epsilon\}, \tag{5}$$

where the *p-values* p^y are defined by

$$p^y := \frac{\left|\{i = 1, \ldots, n \mid \alpha_i^y \leq \alpha_n^y\}\right|}{n}$$

and the conformity scores α_i^y are defined by

$$(\alpha_1^y, \dots, \alpha_n^y) := A\big((x_1, y_1), \dots, (x_{n-1}, y_{n-1}), (x_n, y)\big). \tag{6}$$

Define the *prediction interval* output by CRR as the closure of the convex hull of the prediction set Γ; we will use the notation C_* and C^* for the left and right end-points of this interval, respectively. (Later we will introduce assumptions that will guarantee that Γ itself is an interval from some n on.) As discussed later in Section 5, CRR is computationally efficient: e.g., its computation time is $O(n \ln n)$.

CRR relies on different assumptions about the data as compared with BRR. Instead of the Gaussian model (1), where $\xi_i \sim N(0, \sigma^2)$ and $w \sim N(0, (\sigma^2/a)I)$, it uses the assumption that is standard in machine learning: we consider observations $(x_1, y_1), \dots, (x_n, y_n)$ that are IID (independent and identically distributed).

Proposition 2 (Vovk et al. 2005, Proposition 2.3) *If $(x_1, y_1), \dots, (x_n, y_n)$ are IID observations, the coverage probability of CRR (i.e., the probability of $y_n \in \Gamma$, where Γ is defined by (5)) is at least $1 - \epsilon$.*

Proposition 2 asserts the unconditional validity of CRR. Its validity conditional on the training sequence and the test object is not, however, guaranteed (and it is intuitively clear that ensuring validity conditional on the test object prevents us from relying on the IID assumption about the objects). For a discussion of conditional validity in the context of conformal prediction, see Lei and Wasserman (2014), Section 2, and, more generally, Vovk (2013a). Efficiency (narrowness of the prediction intervals) is not guaranteed either.

The kind of validity asserted in Proposition 2 is sometimes called "conservative validity" since $1-\epsilon$ is only a lower bound on the coverage probability. However, the definition of conformal predictors can be slightly modified (using randomization for treatment of borderline cases) to achieve exact validity; in practice, the difference between conformal predictors and their modified ("smoothed") version is negligible. For details, see, e.g., Vovk et al. (2005), p. 27.

4. Main result

In this section we show that under the Gaussian model (1) complemented by other natural (and standard) assumptions CRR is asymptotically close to BRR, and therefore is approximately conditionally valid and efficient. On the other hand, Proposition 2 guarantees the unconditional validity of CRR under the IID assumption, regardless of whether (1) holds.

In this section we assume an infinite sequence of observations $(x_1, y_1), (x_2, y_2), \dots$ but consider only the first n of them and let $n \to \infty$. We make both the IID assumption about the objects $x_1, x_2, \dots$ (the objects are generated independently from the same distribution) and the assumption (1); however, we relax the assumption that w is distributed as $N(0, (\sigma^2/a)I)$. These are all the assumptions used in our main result:

(A1) The random objects $x_i \in \mathbb{R}^p$, $i = 1, 2, \dots$, are IID.

(A2) The second-moment matrix $\mathbb{E}(x_1 x_1')$ of x_1 exists and is non-singular.

(A3) The random vector $w \in \mathbb{R}^p$ is independent of $x_1, x_2, \dots$.

(A4) The labels $y_1, y_2, \ldots$ are generated by $y_i = w \cdot x_i + \xi_i$, where ξ_i are Gaussian noise variables distributed as $N(0, \sigma^2)$ and independent between themselves, of the objects x_i, and of w.

Notice that the assumptions imply that the random observations (x_i, y_i), $i = 1, 2, \ldots$, are IID given w. It will be clear from the proof that the assumptions can be relaxed further (but we have tried to make them as simple as possible).

Theorem 3 *Under the assumptions (A1)–(A4), the prediction sets output by CRR are intervals from some n on almost surely, and the differences between the upper and lower ends of the prediction intervals for BRR and CRR are asymptotically Gaussian:*

$$\frac{\sqrt{n}}{\sigma}(B^* - C^*) \xrightarrow{\text{law}} N\left(0, \frac{(\epsilon/2)(1-\epsilon/2)}{\phi^2(z_{\epsilon/2})} - \mu'\Sigma^{-1}\mu\right), \tag{7}$$

$$\frac{\sqrt{n}}{\sigma}(B_* - C_*) \xrightarrow{\text{law}} N\left(0, \frac{(\epsilon/2)(1-\epsilon/2)}{\phi^2(z_{\epsilon/2})} - \mu'\Sigma^{-1}\mu\right), \tag{8}$$

where ϕ is the density of the standard normal distribution $N(0,1)$, $\mu := \mathbb{E}(x_1)$ is the expectation of x_1, and $\Sigma := \mathbb{E}(x_1 x_1')$ is the second-moment matrix of x_1.

The theorem will be proved in Section 6, and in the rest of this section we will discuss it. We can see from (7) and (8) that the symmetric difference between the prediction intervals output by BRR and CRR shrinks to 0 as $O(n^{-1/2})$ in Lebesgue measure with high probability.

Let us first see what the typical values of the standard deviation (the square root of the variance) in (7) and (8) are. The second term in the variance does not affect it significantly since $0 \le \mu'\Sigma^{-1}\mu \le 1$. Indeed, denoting the covariance matrix of x_1 by C and using the Sherman–Morrison formula (see, e.g., Henderson and Searle 1981, (3)), we have:

$$\begin{aligned}\mu'\Sigma^{-1}\mu = \mu'(C + \mu\mu')^{-1}\mu &= \mu'\left(C^{-1} - \frac{C^{-1}\mu\mu'C^{-1}}{1 + \mu'C^{-1}\mu}\right)\mu \\ &= \mu'C^{-1}\mu - \frac{(\mu'C^{-1}\mu)^2}{1 + \mu'C^{-1}\mu} = \frac{\mu'C^{-1}\mu}{1 + \mu'C^{-1}\mu} \in [0,1] \end{aligned} \tag{9}$$

(we write $[0,1]$ rather than $(0,1)$ because C is permitted to be singular: see Appendix A for details). The first term, on the other hand, can affect the variance more significantly, and the significant dependence of the variance on ϵ is natural: the accuracy obtained from the Gaussian model is better for small ϵ since it uses all data for estimating the end-points of the prediction interval rather than relying, under the IID model, on the scarcer information provided by observations in the tails of the distribution generating the labels. Figure 1 illustrates the dependence of the standard deviation of the asymptotic distribution on ϵ. The upper line in it corresponds to $\mu'\Sigma^{-1}\mu = 0$ and the lower line corresponds to $\mu'\Sigma^{-1}\mu = 1$. The possible values for the standard deviation lie between the upper and lower lines. The asymptotic behaviour of the standard deviation as $\epsilon \to 0$ is given by

$$\sqrt{\epsilon(1-\epsilon/2)\pi e^{z_{\epsilon/2}^2} - \theta} \sim (-\epsilon \ln \epsilon)^{-1/2} \tag{10}$$

uniformly in $\theta \in [0,1]$.

The assumptions (A1)–(A4) do not involve a, and Theorem 3 continues to hold if we set $a := 0$; this can be checked by going through the proof of Theorem 3 in Section 6. Theorem 3 can thus also

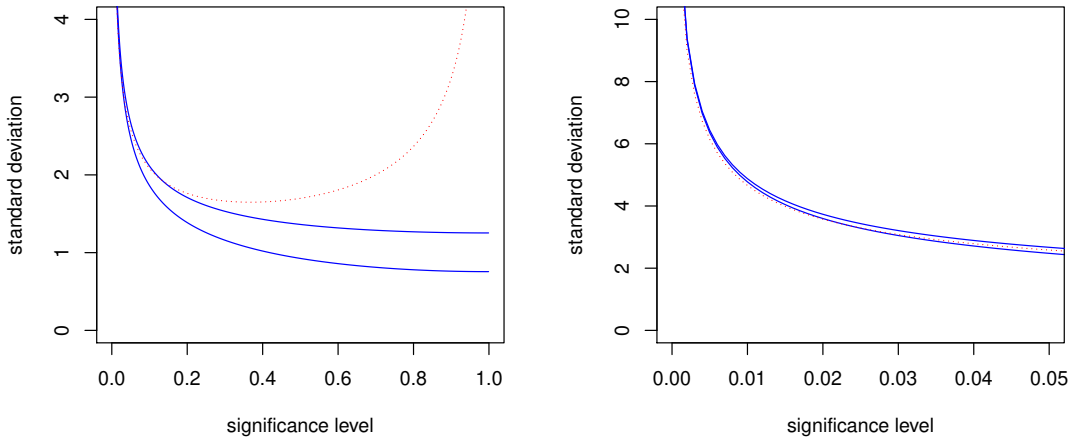

Figure 1: The limits for the standard deviation in Theorem 3 as a function of $\epsilon \in (0,1)$ (left) and $\epsilon \in (0, 0.05]$ (right) shown as solid (blue) lines; the asymptotic expression in (10) shown as a dotted (red) line. In all cases $\sigma = 1$.

be considered as an efficiency result about conformalizing the standard non-Bayesian least squares procedure; this procedure outputs precisely (B_*, B^*) with $a := 0$ as its prediction intervals (see, e.g., Seber and Lee 2003, p. 131). The least squares procedure has guaranteed coverage probability under weaker assumptions than BRR (not requiring assumptions about w); however, its validity is not conditional, similarly to CRR.

5. Further details of CRR

By the definition of the CRR conformity measure, we can rewrite the conformity scores in (6) as

$$\alpha_i^y := \left|\left\{j = 1, \dots, n \mid r_j^y \geq r_i^y\right\}\right| \wedge \left|\left\{j = 1, \dots, n \mid r_j^y \leq r_i^y\right\}\right|, \tag{11}$$

where the vector of residuals $(r_1^y, \dots, r_n^y)'$ is $(I_n - H_n)Y^y$, I_n is the unit $n \times n$ matrix, $H_n := X_n(X_n'X_n + aI)^{-1}X_n'$ is the hat matrix, X_n is the overall design matrix (the $n \times p$ matrix whose ith row is x_i', $i = 1, \dots, n$), and Y^y is the overall vector of labels with the label of the test object set to y (i.e., Y^y is the vector in $\mathbb{R}^n$ whose ith element is y^i, $i = 1, \dots, n-1$, and whose nth element is y). If we modify the definition of CRR replacing (11) by $\alpha_i^y := -r_i^y$, we will obtain the definition of *upper CRR*; and if we replace (11) by $\alpha_i^y := r_i^y$, we will obtain the definition of *lower CRR*. It is easy to see that the prediction set Γ output by CRR at significance level ϵ is the intersection of the prediction sets output by upper and lower CRR at significance levels $\epsilon/2$. We will concentrate on upper CRR in the rest of this paper: lower CRR is analogous, and CRR is determined by upper and lower CRR.

Let us represent the upper CRR prediction set in a more explicit form (following Vovk et al. 2005, Section 2.3). We are given the training sequence $(x_1, y_1), \dots, (x_{n-1}, y_{n-1})$ and a test object x_n; let y be a postulated label for x_n and

$$Y^y := (y_1, \dots, y_{n-1}, y)' = (y_1, \dots, y_{n-1}, 0)' + y(0, \dots, 0, 1)'$$

be the vector of labels. The vector of conformity scores is $-(I_n - H_n)Y^y = -A - yB$, where

$$\begin{aligned} A &:= (I_n - H_n)(y_1, \dots, y_{n-1}, 0)', \\ B &:= (I_n - H_n)(0, \dots, 0, 1)'. \end{aligned}$$

The components of A and B, respectively, will be denoted by $a_1, \dots, a_n$ and $b_1, \dots, b_n$.

If we define

$$S_i := \{y \mid -a_i - b_i y \le -a_n - b_n y\}, \tag{12}$$

the definition of the p-values can be rewritten as

$$p^y := \frac{|\{i = 1, \dots, n \mid y \in S_i\}|}{n};$$

remember that the prediction set is defined by (5). As shown (under a slightly different definition of S_i) in Vovk et al. (2005), pp. 30–34, the prediction set can be computed efficiently, in time $O(n \ln n)$.

6. Proof of Theorem 3

For concreteness, we concentrate on the convergence (7) for the upper ends of the conformal and Bayesian prediction intervals, which we rewrite as

$$\sqrt{n}(B^* - C^*) \xrightarrow{\text{law}} N\left(0, \frac{\alpha(1-\alpha)}{f^2(\zeta_\alpha)} - \sigma^2 \mu' \Sigma^{-1} \mu\right), \tag{13}$$

where $\alpha := 1 - \epsilon/2$ and $\zeta_\alpha := z_{\epsilon/2}\sigma$ is the α-quantile of $N(0, \sigma^2)$. We split the proof into a series of steps.

Regularizing the rays in upper CRR

The upper CRR looks difficult to analyze in general, since the sets (12) may be rays pointing in the opposite directions. Fortunately, the awkward case $b_n \le b_i$ ($i < n$) will be excluded for large n under our assumptions (see Lemma 5 below). The following lemma gives a simple sufficient condition for its absence.

Lemma 4 *Suppose that, for each $c \in \mathbb{R}^p \setminus \{0\}$,*

$$(c \cdot x_n)^2 < \sum_{i=1}^{n-1} (c \cdot x_i)^2 + a \|c\|^2, \tag{14}$$

where $\|\cdot\|$ stands for the Euclidean norm. Then $b_n > b_i$ for all $i = 1, \dots, n-1$.

Intuitively, in the case of a small a, (14) being violated for some $c \neq 0$ means that all $x_1, \dots, x_{n-1}$ lie approximately in the same hyperplane, and x_n is well outside it. The condition (14) can be expressed by saying that the matrix $\sum_{i=1}^{n-1} x_i x_i' - x_n x_n' + aI$ is positive definite.

Proof First we assume $a = 0$ (so that ridge regression becomes least squares); an extension to $a \geq 0$ will be easy. In this case H_n is the projection matrix onto the column space $\mathcal{C} \subseteq \mathbb{R}^n$ of the overall design matrix X_n and $I_n - H_n$ is the projection matrix onto the orthogonal complement $\mathcal{C}^\perp$ of $\mathcal{C}$. We can have $b_n \leq b_i$ for $i < n$ (or even $b_n^2 \leq b_1^2 + \dots + b_{n-1}^2$) only if the angle between $\mathcal{C}^\perp$ and the hyperplane $\mathbb{R}^{n-1} \times \{0\}$ is $45°$ or less; in other words, if the angle between $\mathcal{C}$ and that hyperplane is $45°$ or more; in other words, if there is an element $(c \cdot x_1, \dots, c \cdot x_n)'$ of $\mathcal{C}$ such that its last coordinate is $c \cdot x_n = 1$ and its projection $(c \cdot x_1, \dots, c \cdot x_{n-1})'$ onto the other coordinates has length at most 1.

To reduce the case $a > 0$ to $a = 0$ add the p dummy objects $\sqrt{a} e_i \in \mathbb{R}^p$, $i = 1, \dots, p$, labelled by 0 at the beginning of the training sequence; here $e_1, \dots, e_p$ is the standard basis of $\mathbb{R}^p$. ■

Lemma 5 *The case $b_n \leq b_i$ for $i < n$ is excluded from some n on almost surely under (A1)–(A4).*

Proof We will check that (14) holds from some n on. Let us set, without loss of generality, $a := 0$. Let $\Sigma_l := \frac{1}{l} \sum_{i=1}^{l} x_i x_i'$. Since $\lim_{l \to \infty} \Sigma_l = \Sigma$ a.s.,

$$|\lambda_{\min}(\Sigma_l) - \lambda_{\min}(\Sigma)| \to 0 \quad (l \to \infty) \quad \text{a.s.},$$

where $\lambda_{\min}(\cdot)$ is the smallest eigenvalue of the given matrix. Since $\|x_n\|^2 / n \to 0$ a.s.,

$$\frac{1}{n-1} \sum_{i=1}^{n-1} (c \cdot x_i)^2 = c' \Sigma_{n-1} c \geq \lambda_{\min}(\Sigma_{n-1}) \|c\|^2 > \frac{1}{2} \lambda_{\min}(\Sigma) \|c\|^2 > \frac{\|c\|^2 \|x_n\|^2}{n-1} \geq \frac{(c \cdot x_n)^2}{n-1}$$

for all $c \neq 0$ from some n on. ■

Simplified upper CRR

Let us now find the upper CRR prediction set under the assumption that $b_n > b_i$ for all $i < n$ (cf. Lemmas 4 and 5 above). In this case each set (12) is

$$S_i = (-\infty, t_i], \qquad \text{where } t_i := \frac{a_i - a_n}{b_n - b_i},$$

except for $S_n := \mathbb{R}$; notice that only $t_1, \dots, t_{n-1}$ are defined. The p-value p^y for any potential label y of x_n is

$$p^y = \frac{|\{i = 1, \dots, n \mid y \in S_i\}|}{n} = \frac{|\{i = 1, \dots, n-1 \mid t_i \geq y\}| + 1}{n}.$$

Therefore, the upper CRR prediction set at significance level $\epsilon/2$ is the ray

$$(-\infty, t_{(k_n)}],$$

where $k_n := \lceil (1 - \epsilon/2) n \rceil$ and $t_{(k)} = t_{k:(n-1)}$ stands, as usual, for the kth order statistic of $t_1, \dots, t_{n-1}$.

Proof proper

As before, X stands for the design matrix X_{n-1} based on the first $n-1$ observations. A simple but tedious computation (see Appendix A) gives

$$t_i = \frac{a_i - a_n}{b_n - b_i} = \hat{y}_n + (y_i - \hat{y}_i)\frac{1+g_n}{1+g_i}, \tag{15}$$

where $g_i := x_i'(X'X + aI)^{-1}x_n$ (cf. (3)). The first term in (15) is the centre of the Bayesian prediction interval (4); it does not depend on i. We can see that

$$B^* - C^* = \sqrt{1+g_n}\zeta_\alpha - (1+g_n)V_{(k_n)}, \tag{16}$$

where, as defined earlier (see (13)), $\zeta_\alpha := z_{\epsilon/2}\sigma$, and $V_{(k_n)}$ is the k_nth order statistic in the series

$$V_i := \frac{r_i}{1+g_i} \tag{17}$$

of residuals $r_i := y_i - \hat{y}_i$ adjusted by dividing by $1+g_i$. The behaviour of the order statistics of residuals is well studied: see, e.g., the theorem in Carroll (1978). The presence of $1+g_i$ complicates the situation, and so we first show that g_i is small with high probability.

Lemma 6 *Let $\eta_1, \eta_2, \ldots$ be a sequence of IID random variables with a finite second moment. Then $\max_{i=1,\ldots,n} |\eta_i| = o(n^{1/2})$ in probability (and even almost surely) as $n \to \infty$.*

Proof By the strong law of large numbers the sequence $\frac{1}{n}\sum_{i=1}^n \eta_i^2$ converges a.s. as $n \to \infty$, and so $\eta_n^2/n \to 0$ a.s. This implies that $\max_{i=1,\ldots,n} |\eta_i| = o(n^{1/2})$ a.s. ■

Corollary 7 *Under the conditions of the theorem, $\max_{i=1,\ldots,n} |g_i| = o(n^{-1/2})$ in probability.*

Proof Similarly to the proof of Lemma 5, we have, for almost all sequences $x_1, x_2, \ldots$,

$$\max_{i=1,\ldots,n} |g_i| \le \frac{\|x_n\| \max_{i=1,\ldots,n} \|x_i\|}{\lambda_{\min}(X'X + aI)} < 2\frac{\|x_n\| \max_{i=1,\ldots,n} \|x_i\|}{(n-1)\lambda_{\min}(\Sigma)}$$

from some n on. It remains to combine this with Lemma 6 and the fact that, by Assumption (A1), $\|x_n\|$ is bounded by a constant with high probability. ■

Corollary 8 *Under the conditions of the theorem, $n^{1/2}\left(r_{(k_n)} - V_{(k_n)}\right) \to 0$ in probability.*

Proof Suppose that, on the contrary, there are $\epsilon > 0$ and $\delta > 0$ such that $n^{1/2}\left|r_{(k_n)} - V_{(k_n)}\right| > \epsilon$ with probability at least δ for infinitely many n. Fix such ϵ and δ. Suppose, for concreteness, that, with probability at least δ for infinitely many n, we have $n^{1/2}\left(r_{(k_n)} - V_{(k_n)}\right) > \epsilon$, i.e., $V_{(k_n)} < r_{(k_n)} - \epsilon n^{-1/2}$. The last inequality implies that $V_i < r_{(k_n)} - \epsilon n^{-1/2}$ for at least k_n values of i. By the definition (17) of V_i this in turn implies that $r_i < r_{(k_n)} - \epsilon n^{-1/2} + g_i r_{(k_n)}$ for at least k_n values of i. By Corollary 7, however, the last addend is less than $\epsilon n^{-1/2}$ with probability at

least $1-\delta$ from some n on (the fact that $r_{(k_n)}$ is bounded with high probability follows, e.g., from Lemma 9 below). This implies $r_{(k_n)} < r_{(k_n)}$ with positive probability from some n on, and this contradiction completes the proof. ■

The last (and most important) component of the proof is the following version of the theorem in Carroll (1978), itself a version of the famous Bahadur representation theorem (Bahadur, 1966).

Lemma 9 (Carroll 1978, theorem) *Under the conditions of Theorem 3,*

$$n^{1/2}\left|\left(r_{(k_n)}-\zeta_\alpha\right)-\frac{\alpha-F_n(\zeta_\alpha)}{f(\zeta_\alpha)}+\mu'(\hat{w}_n-w)\right|\to 0 \qquad \text{a.s.}, \tag{18}$$

where F_n is the empirical distribution function of the noise $\xi_1,\dots,\xi_{n-1}$ and $\hat{w}_n := (X'X+aI)^{-1}X'Y$ is the ridge regression estimate of w.

For details of the proof (under our assumptions), see Appendix B.

By (16), Corollary 7, and Slutsky's lemma (see, e.g., van der Vaart 1998, Lemma 2.8), it suffices to prove (13) with the left-hand side replaced by $n^{1/2}(V_{(k_n)}-\zeta_\alpha)$. Moreover, by Corollary 8 and Slutsky's lemma, it suffices to prove (13) with the left-hand side replaced by $n^{1/2}(r_{(k_n)}-\zeta_\alpha)$; this is what we will do.

Lemma 9 holds in the situation where w is a constant vector (the distribution of w is allowed to be degenerate). Let R be a Borel set in $(\mathbb{R}^p)^\infty$ such that (18) holds for all $(x_1,x_2,\dots)\in R$, where the "a.s." is now interpreted as "for almost all sequences $(\xi_1,\xi_2,\dots)$". By Lebesgue's dominated convergence theorem, it suffices to prove (13) with the left-hand side replaced by $n^{1/2}(r_{(k_n)}-\zeta_\alpha)$ for a fixed w and a fixed sequence $(x_1,x_2,\dots)\in R$. Therefore, we fix w and $(x_1,x_2,\dots)\in R$; the only remaining source of randomness is $(\xi_1,\xi_2,\dots)$. Finally, by the definition of the set R, it suffices to prove (13) with the left-hand side replaced by

$$n^{1/2}\frac{\alpha-F_n(\zeta_\alpha)}{f(\zeta_\alpha)}-n^{1/2}\mu'(\hat{w}_n-w). \tag{19}$$

Without loss of generality we will assume that $\frac{1}{n}X_n'X_n\to\Sigma$ as $n\to\infty$ (this extra assumption about R will ensure that Lindeberg's condition is satisfied below).

Since $\mathbb{E}(\alpha-F_n(\zeta_\alpha))=0$ and

$$\operatorname{var}(\alpha-F_n(\zeta_\alpha))=\frac{F(\zeta_\alpha)(1-F(\zeta_\alpha))}{n-1}=\frac{\alpha(1-\alpha)}{n-1},$$

where F is the distribution function of $N(0,\sigma^2)$, we have

$$n^{1/2}\frac{\alpha-F_n(\zeta_\alpha)}{f(\zeta_\alpha)}\xrightarrow{\text{law}}N\left(0,\frac{\alpha(1-\alpha)}{f^2(\zeta_\alpha)}\right)\quad(n\to\infty)$$

by the central limit theorem (in its simplest form).

Since $\hat{w}_n=(X'X+aI)^{-1}X'Y$ is the ridge regression estimate,

$$\mathbb{E}(\hat{w}_n-w)=-a(X'X+aI)^{-1}w=:\Delta_n, \tag{20}$$

$$\operatorname{var}(\hat{w}_n) = \sigma^2 (X'X + aI)^{-1} X'X (X'X + aI)^{-1} =: \Omega_n. \tag{21}$$

Furthermore, for $n \to \infty$

$$n^{1/2}\Delta_n = -n^{-1/2} a \left(\frac{X'X}{n} + \frac{aI}{n} \right)^{-1} w \sim -n^{-1/2} a \Sigma^{-1} w \to 0,$$
$$n\Omega_n = \sigma^2 \left(\frac{X'X}{n} + \frac{aI}{n} \right)^{-1} \frac{X'X}{n} \left(\frac{X'X}{n} + \frac{aI}{n} \right)^{-1} \to \sigma^2 \Sigma^{-1}.$$

This gives

$$n^{1/2}\mu'(\hat{w}_n - w) \xrightarrow{\text{law}} N\left(0, \sigma^2 \mu' \Sigma^{-1} \mu\right) \quad (n \to \infty)$$

(the asymptotic, and even exact, normality is obvious from the formula for $\hat{w}_n$).

Let us now calculate the covariance between the two addends in (19):

$$\begin{aligned}
&\operatorname{cov}\left(n^{1/2} \frac{\alpha - F_n(\zeta_\alpha)}{f(\zeta_\alpha)}, -n^{1/2}\mu'(\hat{w}_n - w) \right) \\
&= \frac{n}{f(\zeta_\alpha)} \operatorname{cov}\left(F_n(\zeta_\alpha) - \alpha, \mu'(\hat{w}_n - w) \right) \\
&= \frac{n}{(n-1)f(\zeta_\alpha)} \sum_{i=1}^{n-1} \operatorname{cov}\left(1_{\{\xi_i \le \zeta_\alpha\}} - \alpha, \mu'(\hat{w}_n - w) \right) \\
&= \frac{n}{(n-1)f(\zeta_\alpha)} \sum_{i=1}^{n-1} \mathbb{E}\Big(\left(1_{\{\xi_i \le \zeta_\alpha\}} - \alpha \right) \mu'(X'X + aI)^{-1} X'\xi \Big),
\end{aligned}$$

where $\xi = (\xi_1, \dots, \xi_{n-1})'$ and the last equality uses the decomposition $\hat{w}_n - w = \Delta_n + (X'X + aI)^{-1}X'\xi$ with the second addend having zero expected value. Since

$$\mathbb{E}\, 1_{\{\xi_i \le \zeta_\alpha\}} \mu'(X'X + aI)^{-1} X'\xi = \sum_{j=1}^{n-1} \mathbb{E}\, 1_{\{\xi_i \le \zeta_\alpha\}} A_j \xi_j = \mu_\alpha A_i,$$

where $A_j := \mu'(X'X + aI)^{-1} x_j$, $j = 1, \dots, n-1$, $\mu_\alpha := \mathbb{E}\, 1_{\{\xi_i \le \zeta_\alpha\}} \xi_i = \int_{-\infty}^{\zeta_\alpha} x f(x) dx$. An easy computation gives $\mu_\alpha = -\sigma^2 f(\zeta_\alpha)$, and so we have

$$\begin{aligned}
\operatorname{cov}\left(n^{1/2} \frac{\alpha - F_n(\zeta_\alpha)}{f(\zeta_\alpha)}, -n^{1/2}\mu'(\hat{w}_n - w) \right) &= \frac{n}{(n-1)f(\zeta_\alpha)} \sum_{i=1}^{n-1} \mu_\alpha A_i \\
&= -\sigma^2 \frac{n}{(n-1)} \sum_{i=1}^{n-1} A_i = -\sigma^2 \mu' \left(\frac{1}{n} X'X + \frac{a}{n} I \right)^{-1} \bar{x} \to -\sigma^2 \mu' \Sigma^{-1} \mu
\end{aligned}$$

as $n \to \infty$, where $\bar{x}$ is the arithmetic mean of $x_1, \dots, x_{n-1}$. Finally, this implies that (19) converges in law to

$$N\left(0, \frac{\alpha(1-\alpha)}{f^2(\zeta_\alpha)} + \sigma^2 \mu' \Sigma^{-1} \mu - 2\sigma^2 \mu' \Sigma^{-1} \mu \right) = N\left(0, \frac{\alpha(1-\alpha)}{f^2(\zeta_\alpha)} - \sigma^2 \mu' \Sigma^{-1} \mu \right);$$

the asymptotic normality of (19) follows from the central limit theorem with Lindeberg's condition, which holds since (19) is a linear combination of the noise random variables $\xi_1, \dots, \xi_{n-1}$ with

coefficients whose maximum is $o(1)$ as $n \to \infty$ (this uses the assumption $\frac{1}{n}X_n'X_n \to \Sigma$ made earlier).

A more intuitive (but not necessarily simpler) proof can be obtained by noticing that $\hat{w}_n - w$ and the residuals are asymptotically (precisely when $a = 0$) independent.

7. Conclusion

The results of this paper are asymptotic; it would be very interesting to obtain their non-asymptotic counterparts. In non-asymptotic settings, however, it is not always true that conformalized ridge regression loses little in efficiency as compared with the Bayesian prediction interval; this is illustrated in Vovk et al. (2005), Section 8.5, and illustrated and explained in Vovk et al. (2009). The main difference is that CRR and Bayesian predictor start producing informative predictions after seeing a different number of observations. CRR, like any other conformal predictor (or any other method whose validity depends only on the IID assumption), starts producing informative predictions only after the number of observations exceeds the inverse significance level $1/\epsilon$. After this theoretical lower bound is exceeded, however, the difference between CRR and Bayesian predictions quickly becomes very small.

Another interesting direction of further research is to extend our results to kernel ridge regression.

Acknowledgements

We are grateful to Albert Shiryaev for inviting us in September 2013 to Kolmogorov's dacha in Komarovka, where this project was conceived, and to Glenn Shafer for his advice about terminology. Thanks to the reviewers for their comments. The first author has been partially supported by the Laboratory for Structural Methods of Data Analysis in Predictive Modeling, Moscow Institute of Physics and Technology, Russian Government grant (agreement 11.G34.31.0073). The second author has been partially supported by EPSRC (grant EP/K033344/1).

References

R. Raj Bahadur. A note on quantiles in large samples. *Annals of Mathematical Statistics*, 37: 577–580, 1966.

Raymond J. Carroll. On the distribution of quantiles of residuals in a linear model. Technical Report Mimeo Series No. 1161, Department of Statistics, University of North Carolina at Chapel Hill, March 1978. Available from `http://www.stat.ncsu.edu/information/library/mimeo.php`.

Samprit Chatterjee and Ali S. Hadi. *Sensitivity Analysis in Linear Regression*. Wiley, New York, 1988.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002.

Harold V. Henderson and Shayle R. Searle. On deriving the inverse of a sum of matrices. *SIAM Review*, 23:53–60, 1981.

Jing Lei and Larry Wasserman. Distribution free prediction bands for nonparametric regression. *Journal of the Royal Statistical Society B*, 76:71–96, 2014.

Ronald H. Randles, Thomas P. Hettmansperger, and George Casella. Introduction to the Special Issue: Nonparametric statistics. *Statistical Science*, 19:561, 2004.

George A. F. Seber and Alan J. Lee. *Linear Regression Analysis*. Wiley, Hoboken, NJ, second edition, 2003.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Machine Learning*, 92: 349–376, 2013a.

Vladimir Vovk. Kernel ridge regression. In Bernhard Schölkopf, Zhiyuan Luo, and Vladimir Vovk, editors, *Empirical Inference: Festschrift in Honour of Vladimir N. Vapnik*, chapter 11, pages 105–116. Springer, Berlin, 2013b.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. On-line predictive linear regression. *Annals of Statistics*, 37:1566–1590, 2009.

Larry Wasserman. Frasian inference. *Statistical Science*, 26:322–325, 2011.

Appendix A. Various computations

For the reader's convenience, this appendix provides details of various routine calculations.

A singular C in (9)

Apply (9) to $\Sigma_\epsilon := \Sigma + \epsilon I$ and $C_\epsilon := C + \epsilon I$, where $\epsilon > 0$, in place of Σ and C, respectively, and let $\epsilon \to 0$.

Computing t_i for simplified upper CRR

In addition to the notation X for the design matrix X_{n-1} based on the first $n-1$ observations, we will use the notation H for the hat matrix $X(X'X + aI)^{-1}X'$ based on the first $n-1$ observations and $\bar{H}$ for the hat matrix $X_n(X_n'X_n + aI)^{-1}X_n'$ based on the first n observations; the elements of H will be denoted as $h_{i,j}$ and the elements of $\bar{H}$ as $\bar{h}_{i,j}$; as always, h_i stands for the diagonal element $h_{i,i}$. To compute t_i we will use the formulas (2.18) in Chatterjee and Hadi (1988).

Since B is the last column of $I_n - H_n$ and

$$\bar{h}_{n,n} = \frac{x_n'(X'X + aI)^{-1}x_n}{1 + x_n'(X'X + aI)^{-1}x_n},$$

we have

$$b_n = 1 - \frac{x_n'(X'X + aI)^{-1}x_n}{1 + x_n'(X'X + aI)^{-1}x_n},$$

$$b_i = \frac{-x_n'(X'X + aI)^{-1}x_i}{1 + x_n'(X'X + aI)^{-1}x_n}.$$

Therefore,

$$b_n - b_i = \frac{1 + x_n'(X'X + aI)^{-1}x_i}{1 + x_n'(X'X + aI)^{-1}x_n}.$$

Next, letting $\hat{y}$ stand for the predictions computed from the first $n-1$ observations,

$$\begin{aligned}
a_i &= \sum_{j=1,\dots,n-1:j\neq i} (-\bar{h}_{i,j}y_j) + (1 - \bar{h}_{i,i})y_i \\
&= y_i - \sum_{j=1}^{n-1} \bar{h}_{i,j}y_j \\
&= y_i - \sum_{j=1}^{n-1} h_{i,j}y_j + \sum_{j=1}^{n-1} \frac{x_i'(X'X + aI)^{-1}x_n x_n'(X'X + aI)^{-1}x_j}{1 + x_n'(X'X + aI)^{-1}x_n} y_j \\
&= y_i - \hat{y}_i + \frac{x_i'(X'X + aI)^{-1}x_n x_n'(X'X + aI)^{-1}X'Y}{1 + x_n'(X'X + aI)^{-1}x_n} \\
&= y_i - \hat{y}_i + \frac{x_i'(X'X + aI)^{-1}x_n \hat{y}_n}{1 + x_n'(X'X + aI)^{-1}x_n}
\end{aligned}$$

for $i < n$, and

$$\begin{aligned}
a_n &= \sum_{j<n} (-\bar{h}_{n,j}y_j) = -\sum_{j=1}^{n-1} \frac{x_j'(X'X + aI)^{-1}x_n}{1 + x_n'(X'X + aI)^{-1}x_n} y_j \\
&= -\frac{Y'X(X'X + aI)^{-1}x_n}{1 + x_n'(X'X + aI)^{-1}x_n}.
\end{aligned}$$

Therefore,

$$a_i - a_n = y_i - \hat{y}_i + \frac{1 + x_i'(X'X + aI)^{-1}x_n}{1 + x_n'(X'X + aI)^{-1}x_n}\hat{y}_n.$$

This gives

$$t_i = (y_i - \hat{y}_i)\frac{1 + x_n'(X'X + aI)^{-1}x_n}{1 + x_i'(X'X + aI)^{-1}x_n} + \hat{y}_n,$$

i.e., (15).

Expressing μ_α via ζ_α

First we use the substitution $y := x^2/2\sigma^2$ to obtain

$$\frac{1}{\sqrt{2\pi}\sigma} \int_0^{\zeta_\alpha} e^{-x^2/2\sigma^2} x dx = \frac{\sigma}{\sqrt{2\pi}} \int_0^{\zeta_\alpha^2/2\sigma^2} e^y dy = \frac{\sigma}{\sqrt{2\pi}} \left(1 - e^{-\zeta_\alpha^2/2\sigma^2}\right). \tag{22}$$

Replacing ζ_α by ∞,

$$\frac{1}{\sqrt{2\pi}\sigma} \int_0^{\infty} e^{-x^2/2\sigma^2} x dx = \frac{\sigma}{\sqrt{2\pi}}. \tag{23}$$

Finally, subtracting (23) from (22) gives

$$\mu_\alpha = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\zeta_\alpha} e^{-x^2/2\sigma^2} x dx = -\frac{\sigma}{\sqrt{2\pi}} e^{-\zeta_\alpha^2/2\sigma^2} = -\sigma^2 f(\zeta_\alpha).$$

Appendix B. Proof of Lemma 9

The proof is modelled on the proof of the theorem in Carroll's technical report (Carroll, 1978) and on Section 2 of Bahadur (1966). We cannot use the result of Carroll (1978) since our conditions are somewhat different. Following Carroll (1978), we only consider the case of simple linear regression ($p = 1$). We will prove that (18) holds for all w, so that w will be a constant vector in $\mathbb{R}^p$ throughout the proof.

We start from the speed of convergence in the ridge regression estimate of regression weights. Let $a_n := n^{-1/2} \ln n$.

Lemma 10 *Under our conditions,* $|\hat{w}_n - w| = o(a_n)$ *a.s.*

Proof This follows immediately from (20) and (21). ∎

The proof uses the following random variables:

$$G_n(x) := n^{-1}\sum_{i=1}^{n}\Big(1_{\{r_i \le x\}} - 1_{\{\xi_i \le \zeta_\alpha\}} - F\big(x + x_i(\hat{w}_n - w)\big) + F(\zeta_\alpha)\Big),$$
$$H_n := n^{1/2}\sup_{x\in J_n}|G_n(x)|,$$

where $J_n := [\zeta_\alpha - a_n, \zeta_\alpha + a_n]$ and

$$W_n(s,t) = n^{-1/2}\sum_{i=1}^{n}\Big(1_{\{\xi_i \le \zeta_\alpha + a_n s + a_n t x_i\}} - 1_{\{\xi_i \le \zeta_\alpha\}} - F(\zeta_\alpha + a_n s + a_n t x_i) + F(\zeta_\alpha)\Big). \quad (24)$$

Lemma 11 *Under our conditions,*

$$\sup\{|W_n(s,t)| \mid s,t \in [0,1]\} \to 0 \quad \textit{a.s.} \quad (25)$$

and, therefore, $H_n \to 0$ *a.s.*

Proof Since $r_i = \xi_i - x_i(\hat{w}_n - w)$ and $\hat{w}_n - w = o(a_n)$ a.s., it is indeed true that (25) implies $H_n \to 0$ a.s.; therefore, we will only prove (25). Let $b_n \sim \ln^2 n$ be a sequence of positive integers. It suffices to consider only s and t of the form $\eta_{r,n} := r/b_n$ for $r = 0, \dots, b_n$. To see this, apply Taylor's expansion: if $|s - \eta_{r,n}| \le b_n^{-1}$ and $|t - \eta_{p,n}| \le b_n^{-1}$, then

$$\left|n^{-1}\sum_{i=1}^{n}\left(F(\zeta_\alpha + sa_n + ta_n x_i) - F(\zeta_\alpha + \eta_{r,n}a_n + \eta_{p,n}a_n x_i)\right)\right|$$
$$\le n^{-1}\sum_{i=1}^{n} f(\zeta^*)a_n b_n^{-1}(1 + |x_i|) = O(a_n b_n^{-1}) = o(n^{-1/2}) \quad \text{a.s.}$$

for some ζ^* (we have used the integrability of x_1).

For fixed s and t we can apply Bernstein's inequality (see, e.g., Györfi et al. 2002, Lemma A.2). Let us fix a sequence $x_1, x_2, \ldots$ such that $\frac{1}{n}\sum_{i=1}^n x_i \to \mu$ (which happens with probability one under our conditions for some μ, namely for $\mu := \mathbb{E}(x_1)$). The cumulative variance (conditional on $x_1, x_2, \ldots$) of the addends in (24) does not exceed

$$\sum_{i=1}^n (a_n s + a_n t x_i) = O(n a_n)$$

a.s. (this again uses the integrability of x_1); therefore, for any $\epsilon > 0$,

$$\mathbb{P}\left\{|W_n(s,t)| > \epsilon\right\} \le c_0 \exp\left(-c_1 n^{1/4}\right)$$

from some n on, where c_0 and c_1 are constants depending on ϵ. The probability that $|W_n(s,t)| > \epsilon$ for some $n \ge N$ and some s, t of the form $\eta_{r,n}$ does not exceed

$$\sum_{n=N}^{\infty} b_n^2 c_0 \exp\left(-c_1 n^{1/4}\right) \to 0 \quad (N \to \infty) \quad \text{a.s.}$$

This completes the proof of the lemma. ■

Remember that $k_n = \lceil \alpha n \rceil$.

Lemma 12 *From some n on, $r_{(k_n)} \in J_n$ a.s.*

Proof We will only show that $r_{(k_n)} \le \zeta_\alpha + a_n$ from some n on a.s. Since

$$\mathbb{P}\left\{r_{(k_n)} > \zeta_\alpha + a_n\right\} \le \mathbb{P}\left\{\sum_{i=1}^n 1_{\{\xi_i \le \zeta_\alpha + a_n + x_i(\hat{w}_n - w)\}} \le k_n\right\}.$$

By Lemma 10, it suffices to show the existence of an $\epsilon > 0$ for which $Q_N(\epsilon) \to 0$ as $N \to \infty$, where

$$\begin{aligned} Q_N(\epsilon) &:= \mathbb{P}\left\{\sum_{i=1}^n 1_{\{\xi_i \le \zeta_\alpha + a_n + t a_n x_i\}} \le k_n \text{ for some } t \in [0,\epsilon] \text{ and } n \ge N\right\} \\ &= \mathbb{P}\Bigg\{F_n(\zeta_\alpha + a_n) \le k_n/n + n^{-1}\sum_{i=1}^n \Big(F(\zeta_\alpha + a_n) - F(\zeta_\alpha + a_n + t a_n x_i)\Big) \\ &\qquad\qquad - n^{-1/2}\left(W_n(1,t) - W_n(1,0)\right) \text{ for some } t \in [0,\epsilon] \text{ and } n \ge N\Bigg\}. \end{aligned}$$

Using Lemma 11 and the fact that

$$n^{-1}\sum_{i=1}^n \Big(F(\zeta_\alpha + a_n) - F(\zeta_\alpha + a_n + t a_n x_i)\Big)$$

$$
\begin{aligned}
&= -n^{-1}\sum_{i=1}^{n} f(\zeta_\alpha + a_n) t a_n x_i + O\left(n^{-1}\sum_{i=1}^{n} t^2 a_n^2 x_i^2\right) \\
&= -n^{-1}\sum_{i=1}^{n} f(\zeta_\alpha) t a_n x_i + O\left(n^{-1}\sum_{i=1}^{n} t a_n^2 x_i\right) + O(a_n^2) \\
&= -f(\zeta_\alpha) t a_n \mu + O\left((\ln\ln n)^{1/2} n^{-1/2} a_n\right) + O(a_n^2) \\
&= -f(\zeta_\alpha) t a_n \mu + o(n^{-1/2}) \quad \text{a.s.}
\end{aligned}
$$

(where $\mu := \mathbb{E}(x_1)$), we obtain

$$
Q_N(\epsilon) = \mathbb{P}\Big\{F_n(\zeta_\alpha + a_n) \le \alpha - t a_n \mu f(\zeta_\alpha) + o(n^{-1/2}) \text{ for some } t \in [0,\epsilon] \text{ and } n \ge N\Big\}.
$$

By Hoeffding's inequality (see, e.g., Györfi et al. 2002, Lemma A.3), when $\delta > 0$ is sufficiently small,

$$
\mathbb{P}\left\{F_n(\zeta_\alpha + a_n) \le \alpha + \delta a_n\right\} \le \exp\left(-cna_n^2\right) = n^{-c\ln n}
$$

for some constant $c > 0$. This implies that indeed $Q_N(\epsilon) \to 0$ as $N \to \infty$. ■

Now we can finish the proof of Lemma 9. Let E_n be the empirical distribution function of r_i. Lemma 10 and the second order Taylor expansion imply

$$
\begin{aligned}
G_n(r_{(k_n)}) &= E_n(r_{(k_n)}) - F_n(\zeta_\alpha) \\
&\quad - n^{-1}\sum_{i=1}^{n}\Big(F(r_{(k_n)}) + f(r_{(k_n)}) x_i(\hat{w}_n - w) - F(\zeta_\alpha)\Big) + O\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) o(a_n^2) \\
&= E_n(r_{(k_n)}) - F_n(\zeta_\alpha) - F(r_{(k_n)}) + F(\zeta_\alpha) + n^{-1}\sum_{i=1}^{n} f(r_{(k_n)}) x_i(\hat{w}_n - w) + o(n^{-1/2}) \quad \text{a.s.}
\end{aligned} \tag{26}
$$

Similarly,

$$
G_n(\zeta_\alpha) = E_n(\zeta_\alpha) - F_n(\zeta_\alpha) - F(\zeta_\alpha) + F(\zeta_\alpha) + n^{-1}\sum_{i=1}^{n} f(\zeta_\alpha) x_i(\hat{w}_n - w) + o(n^{-1/2}) \quad \text{a.s.} \tag{27}
$$

Subtracting (26) from (27) and using Lemmas 11 and 12 and the fact that $E_n(r_{(k_n)}) = k_n/n$, we obtain

$$
\begin{aligned}
&n^{1/2}\left|F(r_{(k_n)}) - F(\zeta_\alpha) - k_n/n + E_n(\zeta_\alpha)\right| \\
&\qquad \le n^{1/2} n^{-1}\sum_{i=1}^{n}\left|f(r_{(k_n)}) - f(\zeta_\alpha)\right| x_i(\hat{w}_n - w)) = o(n^{1/2} a_n^2) \to 0 \quad \text{a.s.}
\end{aligned} \tag{28}
$$

The statement of Lemma 9 can now be obtained by plugging

$$
F(r_{(k_n)}) - F(\zeta_\alpha) = (r_{(k_n)} - \zeta_\alpha) f(\zeta_\alpha) + o(n^{-1/2}) \quad \text{a.s.}
$$

(which follows from the second order Taylor expansion and Lemma 12) and

$$E_n(\zeta_\alpha) = F_n(\zeta_\alpha) + n^{-1/2} W_n(0, a_n^{-1}(\hat{w}_n - w)) + n^{-1} \sum_{i=1}^{n} \Big(F(\zeta_\alpha + x_i(\hat{w}_n - w)) - F(\zeta_\alpha) \Big)$$

(which follows from the definition of W) into (28). Indeed, the addend involving W_n is $o(n^{-1/2})$ a.s. by Lemma 11 and, as we will see momentarily,

$$n^{-1} \sum_{i=1}^{n} \big(F(\zeta_\alpha + x_i(\hat{w}_n - w)) - F(\zeta_\alpha) \big) - f(\zeta_\alpha) \mu (\hat{w}_n - w) = o(n^{-1/2}) \quad \text{a.s.} \tag{29}$$

Therefore, it remains to prove (29). By the second order Taylor expansion, the minuend on the left-hand side of (29) can be rewritten as

$$n^{-1} \sum_{i=1}^{n} f\big(\zeta_\alpha\big) x_i (\hat{w}_n - w)) + O\left(n^{-1} \sum_{i=1}^{n} x_i^2 \right) o(a_n^2)$$
$$= n^{-1} \sum_{i=1}^{n} f(\zeta_\alpha) x_i (\hat{w}_n - w)) + o(n^{-1/2}) \quad \text{a.s.} \tag{30}$$

where we have used $a_n^{-1}(\hat{w}_n - w) \to 0$ a.s. (Lemma 10) and $\mathbb{E}\, x_1^2 < \infty$. And the difference between the first addend of (30) and the subtrahend on the left-hand side of (29) is $O(n^{-1} a_n (n \ln \ln n)^{1/2}) = o(n^{-1/2})$.